View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Keele Research Repository

Received: 22 October 2020 | Accepted: 23 October 2020

DOI: 10.1002/sim.8806

Check for updates

RESEARCH ARTICLE

Statistics in Medicine WILEY

A note on estimating the Cox-Snell R^2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome

Richard D. Riley[1] | **Ben Van Calster**[2,3] | **Gary S. Collins**[4,5]

[1]Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK

[2]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[3]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

[4]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[5]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK

Correspondence
Richard D. Riley, Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire ST5 5BG, UK.
Email: r.riley@keele.ac.uk

Funding information
Cancer Research UK, Grant/Award Number: C49297/A27294; Onderzoeksraad, KU Leuven, Grant/Award Number: C24M/20/064; NIHR Biomedical Research Centre, Oxford; Fonds Wetenschappelijk Onderzoek, Grant/Award Number: G0B4716N

In 2019 we published a pair of articles in *Statistics in Medicine* that describe how to calculate the minimum sample size for developing a multivariable prediction model with a continuous outcome, or with a binary or time-to-event outcome. As for any sample size calculation, the approach requires the user to specify anticipated values for key parameters. In particular, for a prediction model with a binary outcome, the outcome proportion and a conservative estimate for the overall fit of the developed model as measured by the Cox-Snell R^2 (proportion of variance explained) must be specified. This proposal raises the question of how to identify a plausible value for R^2 in advance of model development. Our articles suggest researchers should identify R^2 from closely related models already published in their field. In this letter, we present details on how to derive R^2 using the reported C statistic (AUROC) for such existing prediction models with a binary outcome. The C statistic is commonly reported, and so our approach allows researchers to obtain R^2 for subsequent sample size calculations for new models. Stata and R code is provided, and a small simulation study.

KEYWORDS
clinical prediction model, C statistic (AUROC), R squared, sample size

1 | INTRODUCTION

In 2019 we published a pair of articles in *Statistics in Medicine* that describe how to calculate the minimum sample size for developing a multivariable prediction model with a continuous outcome,[1] or with a binary or time-to-event outcome.[2] These approaches have been implemented in the package *pmsampsize* produced for Stata and R by Ensor et al[3] The required sample size aims to minimize model overfitting and to ensure key parameters (such as the model intercept) are estimated precisely. As for any sample size calculation, the approach requires the user to specify anticipated values for key

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

parameters. In particular, for a logistic regression-based prediction model, the outcome proportion, and a conservative estimate for the overall fit of the developed model as measured by the Cox-Snell R^2 (proportion of variance explained) must be specified.[4,5] For example, to minimize overfitting when developing a logistic regression-based prediction model for a binary outcome, we showed that the sample size (number of participants, n) needed to achieve an expected uniform shrinkage factor of S is,

$$n = \frac{P}{(S-1)\ \ln\left(1 - \frac{R^2_{CS}}{S}\right)},$$

where P is the total number of parameters corresponding to the predictors to be considered for inclusion in the model, S is recommended to be ≥ 0.9 (such that predictor effects must be shrink by $\leq$10%), and R^2_{CS} is a conservative guess at the actual overall fit of the model after model development (ie, the adjusted Cox-Snell R^2_{CS}).

This proposal raises the question of how to identify a plausible value for R^2_{CS} in advance of model development. In most clinical fields, previous prediction models already exist. Indeed, often a new prediction model is developed specifically to update or improve (eg, by adding additional predictors) upon the performance of an existing model. Therefore, our articles suggest researchers should identify R^2_{CS} from closely related models already published in their field,[1,2] and use it to inform the value of R^2_{CS} to use in the sample size calculation for the development of their new model. Extraction of R^2_{CS} is straightforward for prediction models with continuous outcomes, as R^2_{CS} is nearly always reported. For binary and time-to-event outcomes, it is rarely reported, but our article explains how to obtain it from other reported measures including the likelihood ratio statistic along with Nagelkerke's R^2, McFadden's R^2, (for binary outcomes) and Royston's D statistic, O'Quigley's R^2, Royston's R^2, and Royston and Sauerbrei's R^2 (for survival outcomes). A widely reported performance measure is the C statistic, which measures the discrimination performance of a model, and for a binary outcome is equivalent to the area under the receiver operating characteristic curve (AUROC). For time-to-event outcomes, we also discussed how to use the approach of Jinks et al to predict Royston's D (and thus subsequently R^2_{CS}) from a reported C statistic from a survival model such as Cox regression.[6] However, we did not present details on how to derive R^2_{CS} when only the C statistic is reported for a prediction model with a binary outcome—which is often the case. Hence, we now address this to further help researchers to implement our sample size proposal.

2 | OBTAINING R^2_{CS} FROM A REPORTED C STATISTIC FOR A PREDICTION MODEL WITH A BINARY OUTCOME

We consider the scenario where a new prediction model for a binary outcome is being developed for a particular target population. Assume that an article exists that describes the performance of a closely related model (eg, similar outcome and target population), which reports the model's C statistic but not the model's R^2_{CS}. We want to use the reported C statistic to estimate the unreported R^2_{CS}, which is needed to base our sample size calculation on. To do this, we proceed as follows.

First, let $\widehat{p}_i$ denote the existing model's predicted risk of the outcome event for an individual (i) conditional on their values of predictors included in the model. We refer to $\text{logit}(\widehat{p}_i) = \text{LP}_i$ as the linear predictor (LP) values of the existing model. Second, assume LP_i is normally distributed in those with the event and also those without the event, with different means but a common variance. Under these (potentially strong) assumptions, the difference in means of these two normal distributions is a function of the C statistic, as described by various authors elsewhere[7-10]; specifically, the difference in means is $\sqrt{2}\,\Phi^{-1}(C)$, where C is the C statistic, and $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution. Third, we simulate a large dataset of LP_i values based on these two normal distributions, whilst also ensuring the overall outcome proportion matches that assumed for the target population. A logistic regression model can then be fitted to this simulated data, and R^2_{CS} obtained post estimation.

The steps can be outlined more formally as:

i. Simulate a large dataset (eg, one million participants)
ii. Assign an outcome of $\text{Y}_i = 0$ (no event) or $\text{Y}_i = 1$ (event) based on sampling from a Bernoulli (ϕ) distribution, where ϕ is the outcome proportion in the article reporting the existing prediction model
iii. Simulate LP_i values for every participant assuming $\text{LP}_i \sim N(0, 1)$ in the non-events group and $\text{LP}_i \sim N(\mu, 1)$ in the events group, where $\mu = \sqrt{2}\,\Phi^{-1}(C)$

iv. Fit a logistic regression to the simulated data; that is,

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \alpha + \beta \text{LP}_i.$$

This fitted model will have the same C statistic as specified in step (iii). The estimated values of α and β ensure a perfect calibration-in-the-large (= 0) and calibration slope (= 1), respectively, in new data from the same assumed target population.

v. Obtain the R^2_{CS} value for this fitted logistic regression model post estimation, for example, by using the *fitstat* command in Stata or the PseudoR2(model, which="CoxSnell") function in the DescTools package in R. Alternatively, it can be calculated directly using

$$R^2_{CS} = 1 - \exp\left(\frac{-\text{LR}}{n}\right)$$

where n is the number of simulated participants (step i) and LR is the likelihood ratio statistic of the fitted logistic regression model. The obtained R^2_{CS} value can now be used in the sample size calculation for the new prediction model.

Stata and R code are provided in the appendix to implement the approach, and we plan to embed within the *pmsampsize* package. Note that, as discussed in our articles,[2,11] the value of R^2_{CS} depends on the outcome proportion in the target population. Therefore, if the outcome proportion is anticipated to be lower than that reported by the article of the existing model (eg, perhaps because outcomes have since improved), then this could be used in step (ii) (and subsequent sample size calculations) instead.

Where there are a few options for the choice of C statistic (eg, based on multiple validation studies of a previous model), we recommend taking the lowest value, as this is conservative (ie, leads to larger required sample sizes for the new model development study). When using the C statistic reported from a model development study, ideally the C statistic should be adjusted for optimism due to any overfitting. For example, this could be the C statistic after a penalized regression approach has been used; the C statistic after optimism-adjustment based on results from bootstrapping[12]; or based on the C statistic estimated in any independent validation (test) datasets.

3 | A SIMULATION STUDY TO INVESTIGATE THE FIVE-STEP PROCESS WHEN THE ASSUMPTIONS ARE POTENTIALLY INCORRECT

Our five-step approach makes strong assumptions of normality of the existing model's LP, with a common variance for both events and non-events groups. These assumptions are a practical compromise, to help researchers elicit an approximate value for R^2_{CS} in situations where only a reported C statistic, so that they can apply our sample size proposal. Further research might investigate whether they are a good approximation in other situations where the assumptions are invalid. For example, in Figure 1 we show the accuracy of the R^2_{CS} estimate from our five-step process, compared with the actual value (ie, that which would have been observed but is unreported), when the overall LP_i distribution is assumed normal, but the LP_i distribution may not be normal with common variances in the events and non-events groups. We generated 100 different LP_i distributions (ie, 100 different true prediction models) with $\text{LP}_i \sim N(\mu, \sigma^2)$, $\mu \sim$ uniform(0, 5) and $\sigma \sim$ uniform(0.5,3), to cover a range of true C statistic values from about 0.63 to 0.94, corresponding to true R^2_{CS} values of about 0.002 to 0.49. Reassuringly, there is still close agreement between the estimated and actual R^2_{CS} in most scenarios [Figure 1], even though the assumptions made in the five-step process are not necessarily correct.

4 | APPLIED EXAMPLE

Thangaratinam et al[13] developed a prediction model for calculating the risk of an adverse maternal outcome by discharge, in women with early onset preeclampsia in the context of current care. Upon external validation in the target population, the reported C statistic was 0.81. The R^2_{CS} was not provided, and so we applied the five-step procedure described in the previous section, assuming a C statistic of 0.81 and an outcome proportion of 0.77 as reported in the validation study. This gave a R^2_{CS} of 0.21.

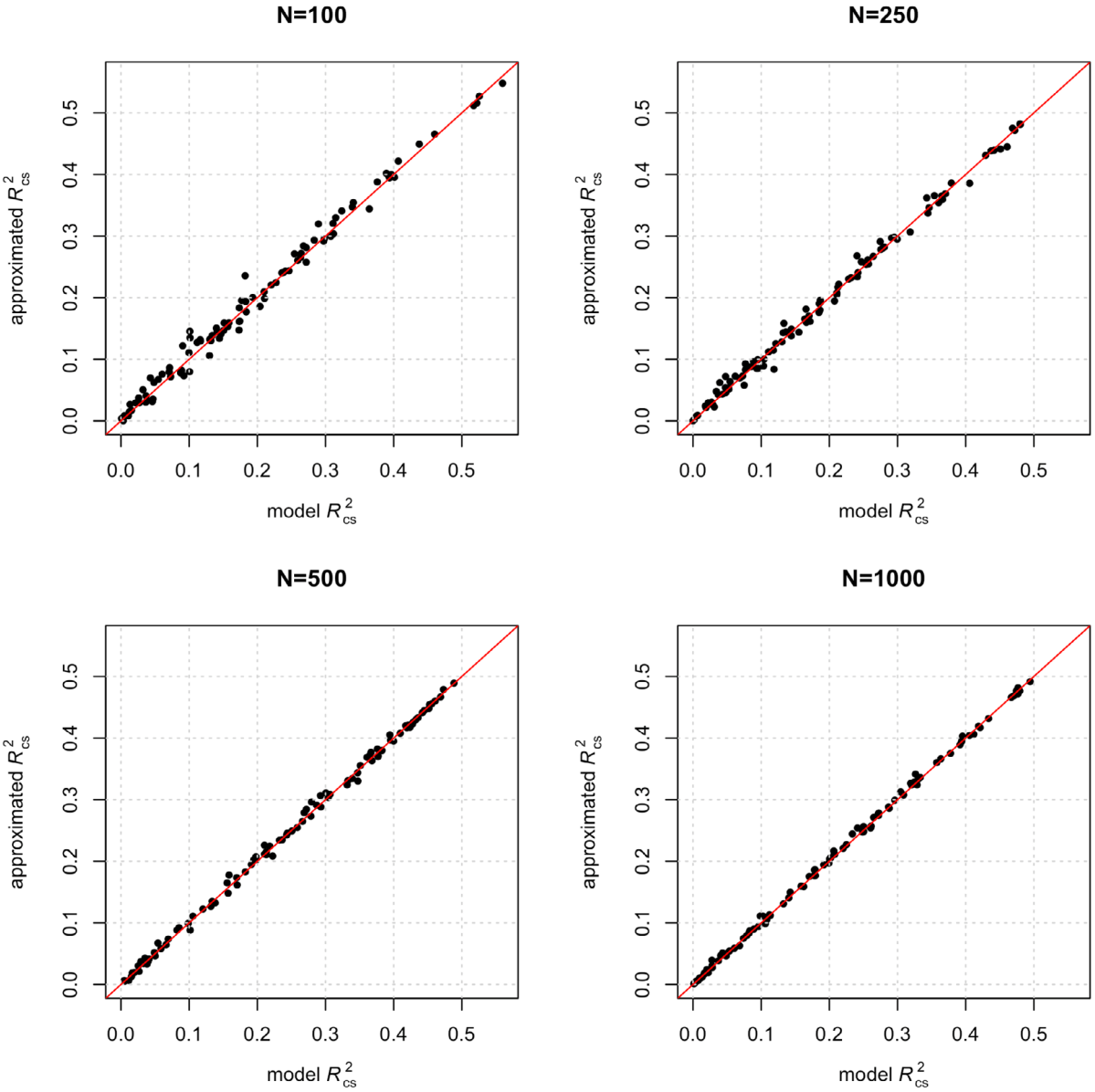

FIGURE 1 Agreement between the estimated R^2_{CS} value (estimated from the reported C statistic estimate using our five-step procedure) and the actual R^2_{CS} (ie, that which would have been observed but is unreported) in 100 prediction model scenarios corresponding to $LP_i \sim N(\mu, \sigma^2)$ and $\mu \sim$ uniform(0, 5) and $\sigma \sim$uniform(0.5,3) [Color figure can be viewed at wileyonlinelibrary.com]

Therefore, to update and extend the model developed by Thangaratinam et al in the same target population, a R^2_{CS} value of 0.21 can be used in the sample size calculations. For example, assuming the new model aimed to consider up to 30 predictor parameters, applying the sample size criteria of Riley et al (eg, in Stata type: pmsampsize, type(b) rsquared(0.21) parameters(30) prevalence(0.77)) gives a minimum sample size required for model development of 1130 participants, with 871 events, and thus an events per predictor parameter of 29.

5 | CONCLUDING REMARK

We have shown how to derive an estimate of the Cox-Snell R^2 from a reported C statistic of a prediction model for a binary outcome. As C statistics (or equivalently AUROCs) are commonly reported for existing prediction models of binary outcomes, our approach allows researchers to quickly obtain a Cox-Snell R^2 to use within sample size calculations when developing new prediction models in the same field.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Richard D. Riley https://orcid.org/0000-0001-8699-0735
Ben Van Calster https://orcid.org/0000-0003-1613-7450
Gary S. Collins https://orcid.org/0000-0002-2772-2316

REFERENCES

1. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. *Stat Med*. 2019;38(7):1262-1275.
2. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.
3. Ensor J, Martin EC, Riley RD. pmsampsize: calculates the minimum sample size required for developing a multivariable prediction model. R Package Version 1.0.3; 2020. https://CRANR-projectorg/package=pmsampsize.
4. Cox DR, Snell EJ. *The Analysis of Binary Data*. 2nd ed. London, UK: Chapman and Hall; 1989.
5. Magee L. R-squared measures based on Wald and Likelihood ratio joint significance tests. *Am Stat*. 1990;44(3):250-253.
6. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol*. 2015;15:82.
7. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
8. Pencina MJ, D'Agostino RB, Massaro JM. Understanding increments in model performance metrics. *Lifetime Data Anal*. 2013;19(2):202-218.
9. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *J Am Stat Assoc*. 1993;88(424):1350-1355.
10. Zelen M, Severo NC. Probability function. In: Abramowitz M, ed. *Handbook of Mathematical Functions*. Washington, DC: National Bureau of Standards Applied Mathematics Series; 1964:925-995.
11. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
12. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
13. Thangaratinam S, Allotey J, Marlin N, et al. Development and validation of prediction models for risks of complications in early-onset pre-eclampsia (PREP): a prospective cohort study. *Health Technol Assess*. 2017;21(18):1-100.

How to cite this article: Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R^2 from a reported *C* statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Statistics in Medicine*. 2020;1–6. https://doi.org/10.1002/sim.8806

APPENDIX A

A1. Stata code to calculate R^2_{CS} from a reported *C* statistic

```
clear all
    * define the existing model's reported C statistic
    local C = 0.81
    * define outcome proportion
    local prev = 0.77

    * define LP distribution
    * events: LP ~ N(0, 1)
    * non-events: LP ~ N(mu, 1)
    * define mu as a function of the C statistic
    local mu = sqrt (2)*(invnorm('C'))

    * now we generate large dataset
    set obs 10000000
    set seed 1234
```

```
* randomly generate outcome proportion according to the outcome proportion
gen outcome = rbinomial(1,'prev')
* specify LP for events and non-events group
* non-events group
gen LP = rnormal(0, 1)
* events group
replace LP = rnormal('mu', 1) if outcome == 1

* Fit a logistic regression with LP as covariate;
* this is essentially a calibration model, and the intercept and slope estimates
* will ensure the outcome proportion is accounted for, without changing C statistic
logistic outcome LP, coef
fitstat
* report Cox-Snell R-squared
disp r(r2_ml)
* gives R2 of 0.21 (correspond to an R2 Nagelkerke of 0.32)
```

A2. R code to calculate R^2_{CS} from a reported *C* statistic

```
approximate_R2 <- function(auc, prev, n = 1000000){

 # define mu as a function of the C statistic
 mu <- sqrt(2) * qnorm(auc)

 # simulate large sample linear prediction based on two normals
 # for non-eventsN(0, 1), events and N(mu, 1)

 LP <- c(rnorm(prev*n, mean=0, sd=1), rnorm((1-prev)*n, mean=mu, sd=1))
 y <- c(rep(0, prev*n), rep(1, (1-prev)*n))

 # Fit a logistic regression with LP as covariate;
 # this is essentially a calibration model, and the intercept and
 # slope estimate will ensure the outcome proportion is accounted
 # for, without changing C statistic

 fit <- lrm(y~LP)

 max_R2 <- function(prev){
   1-(prev^prev*(1-prev)^(1-prev))^2
 }
 return(list(R2.nagelkerke = as.numeric(fit$stats['R2']),
             R2.coxsnell = as.numeric(fit$stats['R2']) * max_R2(prev)))
}

> set.seed(1234)
> approximate_R2(auc = 0.81, prev = 0.77, n=1000000)
$R2.nagelkerke
[1] 0.3183689

$R2.coxsnell
[1] 0.2100957
```